DOCUMENT RESUME

ED 286 885 TM 870 418

AUTHOR Tomsic, Margie L.; And Others
TITLE The Effect of Poor Fitting Items on the Distributions of Extended Caution Indices.
PUB DATE Apr 87
NOTE 29p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Achievement Tests; Elementary Education; *Goodness of Fit; *Latent Trait Theory; *Statistical Distributions; Test Construction; Testing Problems; Test Interpretation; Test Items; Test Reliability
IDENTIFIERS Caution Indices; Comprehensive Tests of Basic Skills; *Confidence Intervals (Statistics); *Extended Caution Index; Rasch Model

ABSTRACT

Extended caution indices (ECI) specify the degree of confidence that can be placed in an individual's test score by analyzing patterns of item response. Among the most promising of such indices are the standardized ECIs. Contrary to the literature, several instances were found, in a previous study, of nonnormal distributions of ECIs with samples taken from four different grade levels. Several items from the study failed to fit the Rasch model. The current study examined whether ECI distributions approach normality if worst fitting items are dropped from analyses. The sample consisted of Comprehensive Tests of Basic Skills (CTBS) data from 700 third grade students and 737 seventh grade students. Seven out of eight distributions deviated from normality. Removal of the worst fitting items did not return the distributions to normal. The percentage of students at the extreme ends of the curves remained constant despite the inclusion or deletion of test items. Distributions were most stable at upper grade levels. Given the robustness of these indices, they may still prove useful in identifying mismeasured students. The results of this study posed more questions than answers. Further reesearch is needed, and replication using different measures of item misfit is recommended. (Author/MDE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED286885

THE EFFECT OF POOR FITTING ITEMS ON THE DISTRIBUTIONS OF EXTENDED CAUTION INDICES

Margie L.Tomsic
University of Illinois College of Medicine at Peoria

Stephen J. Schellenberg
Saint Paul Public Schools

Arthur Mittman
University of Oregon

Paper presented at the 1987 Annual Meeting of the American Educational Research Association, Washington, D.C.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
M.L. Tomsic
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM 870 418

BEST COPY AVAILABLE

ABSTRACT

Extended caution indices specify the degree of confidence we can place in an individual's test score by analyzing patterns of item response. Among the most promising of such indices are the standardized ECIs. Contrary to the literature, Tomsic (1986) found several instances of nonnormal distributions of ECIs with samples taken from four different grade levels. Several items from the study failed to fit the Rasch model. The current study examined whether ECI distributions approach normality if worst fitting items are dropped from analyses.

The sample consisted of CTBS test data taken from 700 third grade students and 737 seventh grade students. Seven out of eight distributions deviated from normality. Removal of the worst fitting items did not return the distributions to the normal. The percentage of students at the extreme ends of the curves remained constant despite the inclusion or deletion of test items. Distributions were most stable at upper grade levels. Given the robustness of these indices, they may still prove useful in identifying mismeasured students.

Since the advent of mental testing, theorists have questioned the representativeness of total test scores in measuring student ability levels. Poor test taking strategies and individual variations in test anxiety, language fluency, motivation or background experiences can result in unreliable total scores (Harnisch & Linn, 1981). Examinee behaviors such as cheating or guessing can further complicate the issue (Hulin, Drasgow, & Parsons, 1983; Nelson, 1986; Wright, 1977). Thus, researchers have tried to develop indices that identify mismeasured students.

At least eleven different indices use summary statistics to detect atypical response patterns. Standard summary statistics are based on patterns of right and wrong answers. Of these, those developed most recently are Sato's caution index and the modified caution index (Harnisch & Linn, 1981). Indices of this type organize test information into a binary, student by item matrix called the Student-Problem or S-P curve, where the rows of the matrix represent students while the columns represent items. Students are ranked in descending order according to their total test scores while items are arranged in ascending order from left to right according to their levels of difficulty. We assume that responses are based on something other than ability when extreme deviations from a Guttman scaling pattern are present. Thus, we obtain a high caution index (Tatsuoka & Linn, 1983).

A second category of indices are based on item response theory (IRT). In contrast to indices described earlier, these coefficients no longer depend on item difficulty orders for particular samples of students. Therefore, if we assume that our data fit the IRT model of interest, these person-fit measures are invariant across different samples of items and examinees. Approximately nine different indices appear in the literature. Those named appropriateness measures use the maximum likelihood function to estimate item and ability parameters and can be used for all types of IRT models. Appropriateness measures are described by Drasgow (1982) and by Levine and Rubin (1979). A second type make use of standardized residuals to calculate the weighted or unweighted total fit mean square. The weighted total fit mean square (Rudner, 1983) is the Rasch person-fit statistic (RX^2) (Schmitt & Crocker, 1984). The total fit mean square is based on the one parameter IRT model and thus does not take guessing into account. This statistic has often been used to test item fit to the Rasch model. Limitations and difficulties to consider when using this index are summarized in Hambleton and Murray (1983) and Divgi (1986).

A third category of indices can be viewed as mathematical extensions which connect Sato's index to an IRT model. When these indices are applied, ideal Guttman response patterns are replaced by person response curves, derived theoretically from IRT models. At a given ability level, a person response curve is obtained by varying item difficulty. Item difficulty

thereby becomes a continuous variable defined by a logistic probability function (Harnisch & Tatsuoka, 1983). Two of the more popular indices of this type include Extended Caution Index Two (ECI2) and Extended Caution Index Four (ECI4). ECI2 compares the similarity of the group response curve with observed response and probability vectors. This index is more sensitive to students whose response patterns deviate from others in their ability group. In contrast, ECI4 compares the similarity of the probability curve with observed response and group response vectors. Values of these indices suggest that they are more sensitive to students whose response patterns deviate from their own expectancies (Tatsuoka, 1984).

The conditional variances of these ECI's have a "U" shape (Tatsuoka, 1984). Thus, their variances are inflated at extremely low or high levels of ability. The unstandardized indices have therefore been standardized by subtracting the conditional expectation of each ECI from the original ECI and dividing by the square root of its conditional variance. The standardized forms ofECI2 and ECI4 are defined below in Equations (1) and (2) respectively as

$$ECI2_{iz} = \frac{n \, \mathrm{cov}(\underset{\sim}{P}_i - \underset{\sim}{y}_i, \underset{\sim}{G})}{\left[\sum_{j=1}^{n} (G_j - \bar{G}_\cdot)^2 \sigma_{ij}^2 \right]^{1/2}} \qquad (1)$$

$$ECI4_{iz} = \frac{n \, \mathrm{cov}(P_i - y_i, P_i)}{\left[\sum_{j=1}^{n} (P_{ij} - T_i)^2 \sigma^2_{ij} \right]^{1/2}} \qquad (2)$$

where $y_i = (y_{i1}, \ldots, y_{in})$ is the binary-scored observed response vector for i;
$y_\cdot = (y_{\cdot 1}, \ldots, y_{\cdot n})$ and y_j is the mean score of the item j;
$P_i = (P_{i1}, \ldots, P_{in})$ and P_{ij} is the probability associated with examinee i answering item j correctly and it is obtained from an item characteristic curve

$$G_j = \frac{1}{N} \sum_{i=1}^{N} P_{ij}$$

and $$P_{\cdot\cdot} = \frac{1}{n} \sum_{j=1}^{n} y_{\cdot j} \, .$$

To date, a number of investigations which use data from the National Assessment of Educational Progress (NAEP) suggest that the standardized ECIs hold more promise for detecting students with aberrant response patterns than other indices of this type. Reasons cited for their preference include weak linear and curvilinear relationships with total test scores, increased precision for large samples of data, the ability to substitute Sato's index for ECIs when sample sizes are small, and the approximation of these indices to a normal curve (Birenbaum, 1985; Harnisch & Tatsuoka, 1983; Rudner, 1983; Tatsuoka, 1984).

Some of these findings were replicated across four grade levels by Tomsic and Mittman (1986) using a different sample of test data. In consecutive years, correlations between these

indices were significant but low at upper grade levels and at the chance level at lower grades. Stability was no different for boys and girls or for various sample sizes.

Contrary to the literature, Tomsic (1986) further found that most of the standardized ECI distributions were not perfectly symmetrical, but contained slightly more students in the right hand tail of the distribution for two different CTBS subscales taken from samples of students at four grade levels (Tomsic, 1986). However, depending on one's selection criteria, several items from this same study failed to fit the Rasch model. Therefore, the present investigation examines whether or not nonsymmetrical ECI distributions improve in normality when worst fitting items are dropped from analyses.

Rationale and Research Questions

Numerous person-fit indices appear in the literature, yet, investigators have only begun to examine their statistical properties. Few researchers have examined the sampling distributions of person-fit indices. Studies which do exist are often based solely on a chi-square statistic or are confined to a particular group of students (Traub & Lam, 1985). If one does not know the sampling distribution of an index, one can not apply an accurate statistical test to determine if a coefficient deviates significantly from the normal population. Therefore, an arbitrary criterion is usually set in hopes of accurately identifying students with

atypical test results. Since many statistical tests assume that data follow a normal curve, studies on normality and index distributions are worthy of investigation.

Previous research suggests that standardized ECIs are normally distributed (Tatsuoka & Tatsuoka, 1982; Harnisch & Tatsuoka, 1983). However, in a recent study, Tomsic (1986) found that her ECI distributions showed a slight, positive skew for three out of four grade levels sampled. A greater percentage of students had ECIs more than two standard deviations above the mean and fewer had ECIs less than two standard deviations below the mean than would appear in the normal distribution.

Unlike prior studies in the literature, Tomsic (1986) failed to discard items that deviated from the one parameter model. In one instance, between thirteen and fifteen percent of her items deviated from the Rasch model when using a fit mean square criterion of 1.31. Given the sensitivity of this index to variations in sample size (Hambleton & Murray, 1983), this more stringent criterion was used as an arbitrary cutting point. Regardless of the criterion used, current wisdom suggests that advantages gained from the use of IRT measures are lost when data do not fit the model of interest (Hambleton & Murray, 1983). Therefore, the present authors postulated that if ECI distributions rapidly approach normality when items of worst fit are dropped from analyses, one can conclude that misfitting items seriously effect

the distributions of ECIs. In contrast, failure to find this result might lead one to question the assumption of normality for different samples of ECIs.

Specific questions addressed by this study include:

1) Do the distributions of standardized ECIs continue to deviate from normality when a different sample of students is drawn from the same population of test takers ?

2) Will nonnormal distributions tend towardnormality when items of worst fit are dropped from the test ?

Method

Subjects

Students from a school district in the Pacific Northwest region of the United States supplied data for the study. Test data resulted from the 1984 spring test cycle. During this year, the Comprehensive Tests of Basic Skills was administered to the majority of the district's 17,000 students. In general, the mean score for the district fell above the 70th percentile on the publisher's national norms.

The current investigation used item response data from a random sample of 700 third grade students and 737 seventh grade students. Analyses in the study used data from the Vocabulary and Math Computation subtests. Due to the nature of their content, these subtests were thought originally to be more unidimensional than other subtests contained within the CTBS. The present study used these same subtests in an attempt to delineate the previous findings of Tomsic (1986).

The Test

The CTBS are a series of norm-referenced tests designed to measure school achievement in the basic skills. The Vocabulary subtest consists of thirty items at the third grade level and forty-five items at the seventh grade level. The Math Computation portion of the CTBS contains thirty-six items at the third grade level and forty items at the seventh grade level. The CTBS has two alternate forms and multiple levels. All students in the study took Form U of the test. Those who were third graders took level E, while those who were seventh graders took level H.

The Program

The program used in this study to calculate mean square fit and ECI measures was devised by Nelson and Chatman (1984). This program is written in the the Statistical Application Systems (SAS) programming language PROC MATRIX. In its original

form this program accepts only a data matrix of integers and then scores a test. Since data from the study were prescored, the scoring portion of the program was eliminated. The remaining steps in the program were left intact.

Procedure

The sample consisted of test data from 700 third grade students and 737 seventh grade students, drawn randomly from a 1984 archive tape. Item responses were recoded into a series of ones and zeros, with "1" meaning correct and "0" meaning incorrect. Descriptive statistics were analyzed at each grade level for the two CTBS subtests.

The SAS program written by Chatman and Nelson (1984) produced the standardized versions of ECI2 and ECI4, as well as the item fit mean square statistic for the two subtests at each grade level. Goodness-of-Fit tests, computed with each successive half standard deviation unit, tested the eight ECI distributions against proportions found under a normal curve.

The item fit mean square index showed items of least fit to the Rasch model. This index was chosen as a measure of item fit in place of ECI item fit values, since it correlates highly with the latter (Nelson, 1986), but is derived independently of them. For each distribution, the highest item fit mean square index was dropped from the analysis. A new

run of the SAS program produced standardized ECI values, and Goodness-of-Fit indices were recomputed for the eight distributions. This method of eliminating the worst fitting items was continued until either the chi-square indices failed to deviate from normality or until no significant change in the shape of the the distribution occurred after the five worst fitting items were dropped.

Results

Table 1 shows the means and standard deviations for each grade level and subtest in the study. As expected, students performed well on each exam when compared to the national average. Total test score distributions were not perfectly normal, showing a negative skew.

Insert Table 1 About Here

Initially, percentages from seven out of eight distributions of ECIs deviated significantly from the normal. With the exception of ECI2z for third grade math, this is as expected, since it replicates the findings of Tomsic (1986) using a different sample from the same data set. The chi-square values fell within the same range as those in the earlier sample, as shown in Table 2.

Insert Table 2 About Here

Values for the items of worst fit to the Rasch model are listed in Table 3 by grade and subtest. As we can see, the most extreme items did not deviate beyond a fit mean square index of 2.03. In many analyses, this value would not be considered as extremely deviant. Yet, research suggests that when more stringent or powerful tests of fit are used, more items deviate from the Rasch model than one might expect when using an unadjusted total fit mean square index (Divgi, 1987). This index is further subject to variations in sample size and to test score outliers (Hambleton and Murray,1983).

Insert Table 3 About Here

Even with a more stringent criterion, most distributions deviated from a perfect normal curve. Therefore, the five worst fitting items were removed successively from analyses to determine if normality would result after setting an even more stringent criterion. In most cases, removal of the five items of least fit equalled a cutting score of 1.31, which is the 25th percentile of a chi-square distribution with one degree of freedom.

Table 4 shows chi-square tests for ECI distributions when one through five items are removed from a test. Contrary to predictions, removal of the worst fitting items did not move the distributions toward the normal. In some instances, a nonnormal curve was made more deviant. Chi-square values did not change in any systematic way with the elimination of particular groups of items. For exploratory purposes, more items were dropped from the analyses than depicted in the table with no substantive change in pattern.

Insert Table 4 About Here

Fluctuations in chi-square values were very prominent across the third grade verbal subtest for ECI4z. As items were dropped, the the curve became more erratic in appearance. Of interest is that erratic results became more extreme in the subtest which had the fewest number of items.

Figures 1 through 4 compare the ECIz distributions for each grade level in successive half standard deviation units to the normal curve. As a result of these findings, the authors returned to the one distribution which did not differ from the normal curve (ECI2z for Third Grade Math) and conducted similar analyses. The results differed from that of a normal curve.

Insert Figures 1 through 4 About Here

Despite fluctuations in chi-square values, the removal lines tended towards a single shape when the curves were smoothed by a moving averages technique. In the upper grades, a constant positive skew was particularly noticeable when items were removed successively in a stepwise fashion. This is depicted nicely in Figures 5 and 6.

Insert Figures 5 and 6 About Here

The percentage of students falling two standard deviations above the mean remained constant within a given grade level. For the upper grades, the number of students stayed between four and five percent across differences in the items removed. This stability is shown in Tables 5 and 6.

Insert Tables 5 and 6 About Here.

Discussion

The removal of moderately misfitting test items did not return the ECIz distributions to normality. Although actual chi-square values fluctuated, the shape of the removal lines

remained constant. This suggests that the distributions of ECIzs may be fairly robust to some shortening of tests or to the inclusion of some nonfitting items in the test pool.

For the second time, the authors found that seven out of eight distributions in the study deviated from proportions found under a normal curve when Goodness-of-Fit tests were applied. This suggests that ECIz distributions may not be perfectly symmetrical across all samples and subtests.

Several explanations can be advanced for the nonnormal distributions. Contrary to the literature, ECIz distributions may not be truly normal. An equally viable explanation suggests that something inherent in the sample lead to the nonnormality in ECIz distributions. For instance, many students in this sample had above average test scores. Although total test scores are only weakly correlated with ECIs (Tomsic, 1986; Harnisch & Tatsuoka, 1983), these two phenomena may stem from a common source, namely, well developed test taking skills. Replications of this study on normally distributed samples of data from other school districts could test this possibility.

Regardless of their distributions, these indices seem robust to the inclusion and subsequent deletion of items within a moderate range of fit. This robustness is seen particularly at the extreme ends of the distributions. If extreme values remain remain stable despite the deletion of test items, the indices

may consistently identify the same students with changes in the test item pool. Certainly, we hope that such a result exists if these indices are reliable.

Given the difficulties inherent in relying solely on a statistical test to identify mismeasured students, the quest for the perfect test statistic may be ill advised. The question, then, becomes a matter of locating the optimal cutting criterion which maximizes identification and minimizes false alarms. The best procedure may involve flagging a certain percentage of students at either extremes of the distribution. Ultimately, however, results should be based on multiple sources of information. If person-fit indices are to be of any real practical use in the classroom, they should be validated on teacher observations.

Implications of Future Research

Results from the present study pose more questions than answers. Therefore, further research is needed. The same study might best be replicated on other data sets known to behave in a normal manner. Second, the present study did not include items with fit mean square indices above 2.03. Therefore, a simulated study with more deviant test items is in order to discern whether or not the same results hold true for extremely deviant test items. Finally, the same study should be replicated using several different measures of item misfit.

TABLE 1. Means and Standard Deviations for Third and Seventh Grade Students on Math and Verbal Subscales from the CTBS

	Reading Vocabulary			Math Computation		
	Mean	S.D.	#Items	Mean	S.D.	#Items
Grade 3	25.0	4.5	30	29.0	4.8	36
Grade 7	34.0	7.8	45	23.6	7.0	40

TABLE2. Chi-Square Goodness-of-Fit Statistics for ECI2z When Compared To A Normal Distribution

	Subscales	
	Vocabulary	Math Computation
1986		
Grade 3	50.3 *** df= 11	30.13 ** df= 11
Grade 7	74.6 *** df= 11	72.0 *** df= 11
1987		
Grade 3	47.0 *** df= 11	15.33 df= 11
Grade 7	60.9 *** df= 11	71.70 *** df= 11

** p <.05
*** p <.01

TABLE3. Item Fit Mean Square Statistics for the Misfitting Items

	Reading Vocabulary	Math Computation
Grade 3	2.09 Nitems= 30	1.71 Nitems= 36
Grade 7	1.76 Nitems= 45	1.72 Nitems= 40

TABLE4. Goodness-of-Fit Statistics for Successive Removal of Five Items of Worst Fit

	Reading Vocabulary		Math Computation	
	Grade3	Grade7	Grade3	Grade7
ECI2z				
0 Items Removed:	47.04**	60.93 **	15.33	71.69**
1 Item Removed:	64.02**	56.00**		71.00**
2 Items Removed:	67.19**	67.50**		49.56**
3 Items Removed:	47.45**	109.88**		60.58**
4 Items Removed:	41.62**	67.11**		65.38**
5 Items Removed:	-	79.70**		68.31**
ECI4z				
0 Items Removed:	72.81**	39.57**	28.62**	67.84**
1 Item Removed:	75.29**	33.33**	30.97**	44.20**
2 Items Removed:	48.54**	39.42**	36.03**	50.73**
3 Items Removed:	147.88**	59.20**	29.72**	54.84**
4 Items Removed:	166.27**	38.64**	26.00**	52.18**
5 Items Removed:		49.30**	29.77**	62.68**

* $p < .05$
** $p < .01$

TABLE5. Percentage of Third Grade Students Two or More Standard Deviations Above the Mean With Four Items Removed

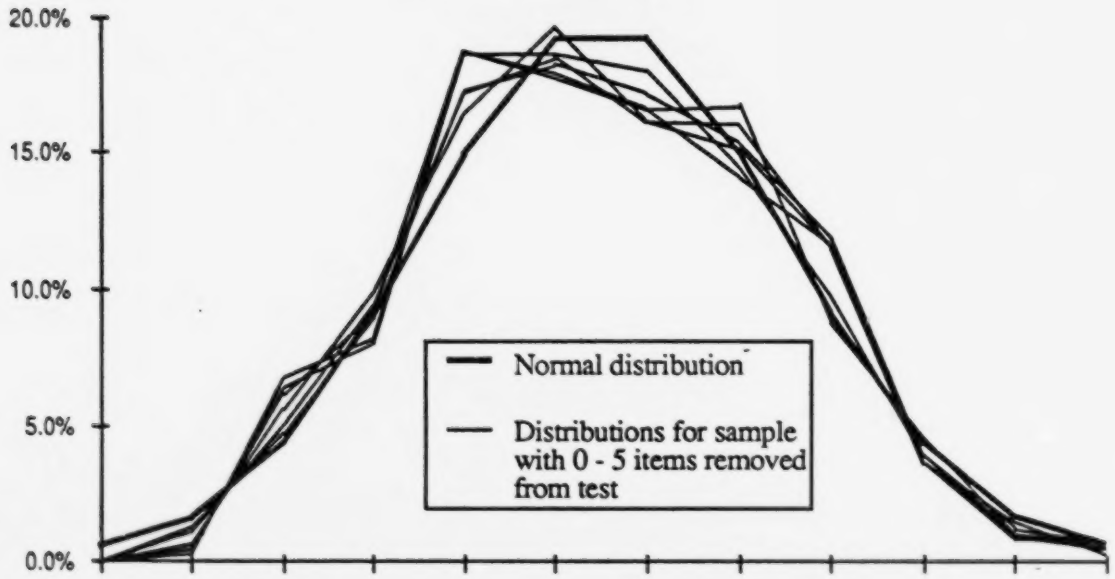
	Reading Vocabulary		Math Computation	
	ECI2z	ECI4z	ECI2z	ECI4z
0 Items Removed:	3.0	2.0	2.0	1.6
1 Item Removed:	3.5	2.0	2.8	2.0
2 Items Removed:	3.5	2.6	2.0	2.0
3 Items Removed:	2.9	2.1	3.5	1.8
4 Items Removed:	2.5	1.7	2.3	1.7
5 Items Removed:				

TABLE6. Percentage of Seventh Grade Students Two or More Standard Deviations Above the Mean With Five Items Removed

	Reading Vocabulary		Math Computation	
	ECI2z	ECI4z	ECI2z	ECI4z
0 Items Removed:	4.0	3.8	5.0	5.0
1 Item Removed:	4.0	4.4	5.1	4.0
2 Items Removed:	4.5	4.2	5.1	4.0
3 Items Removed:	4.6	4.9	5.1	4.0
4 Items Removed:	4.5	4.0	4.9	4.3
5 Items Removed:	4.9	4.7	4.8	4.3

FIGURE 1. DISTRIBUTIONS OF ECI2z FOR THIRD-GRADE TESTS

MATH COMPUTATION



READING VOCABULARY

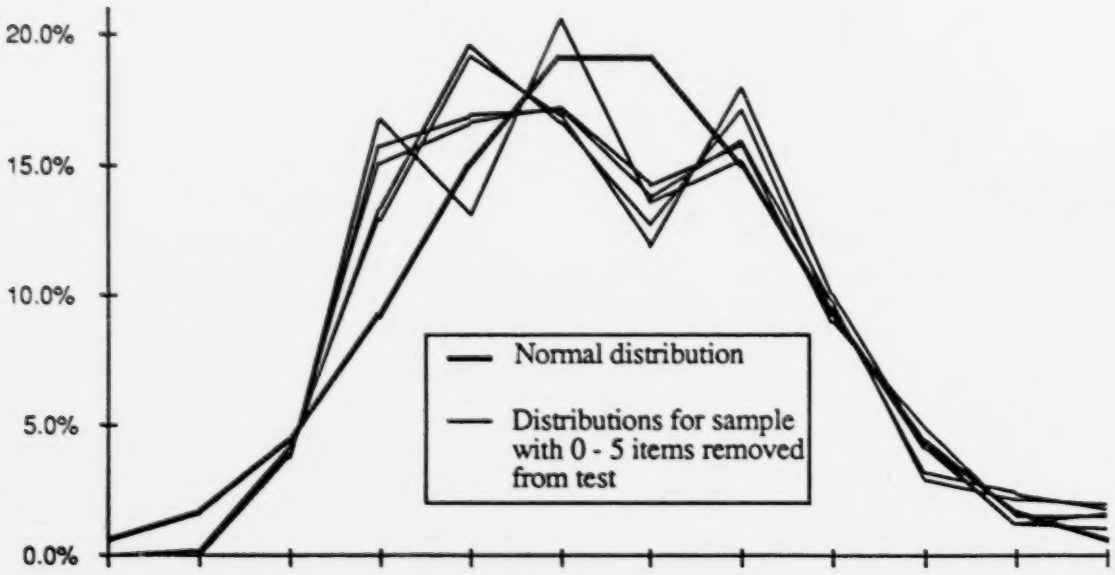
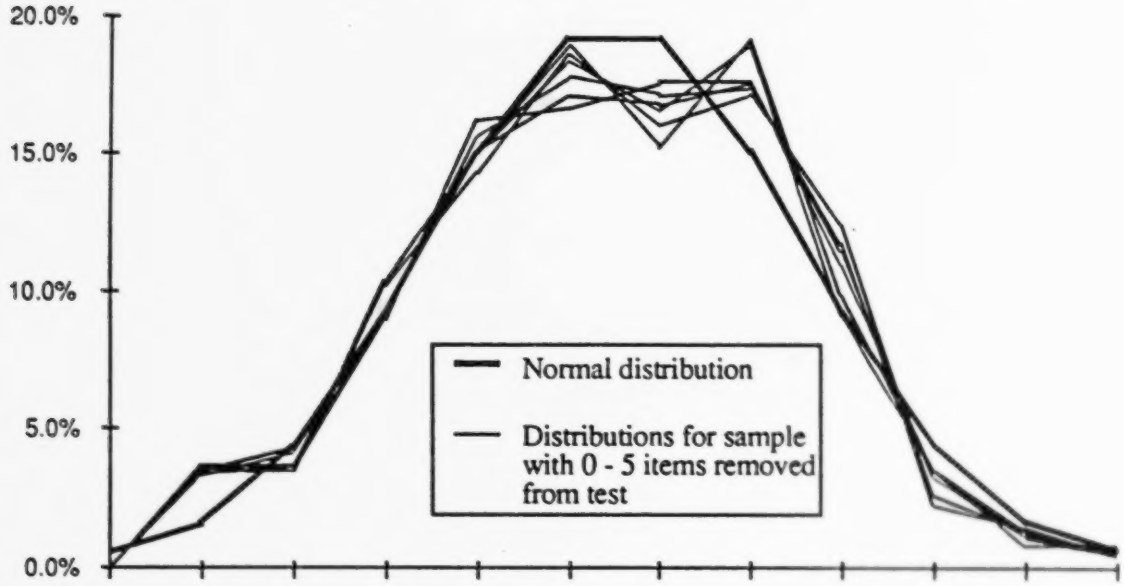


FIGURE 2. DISTRIBUTIONS OF EC14z FOR THIRD-GRADE TESTS

MATH COMPUTATION



READING VOCABULARY

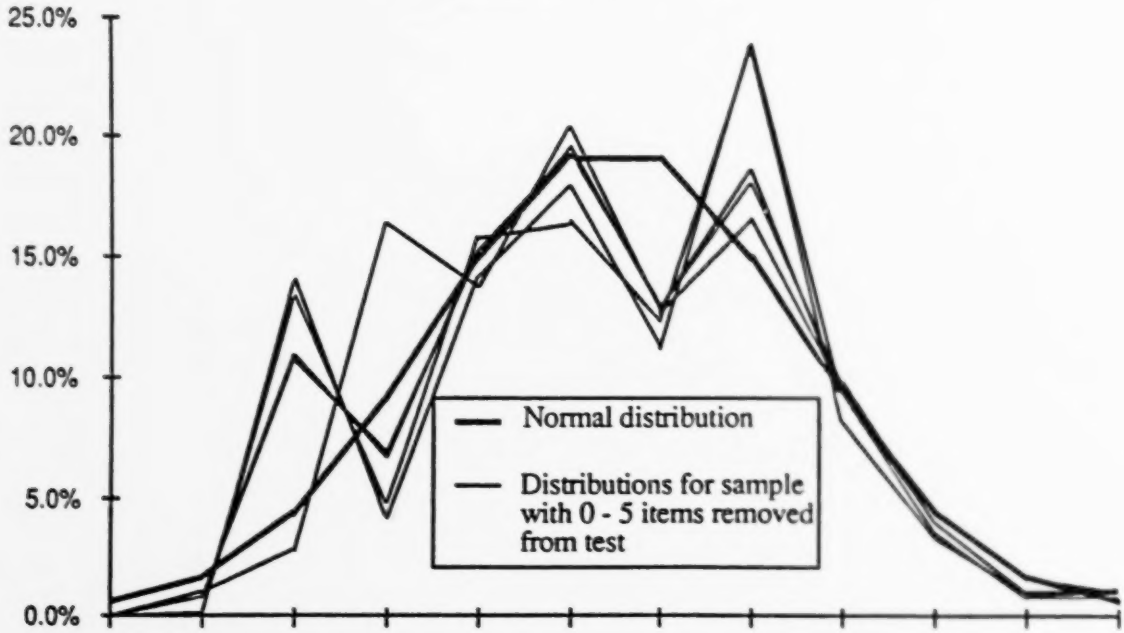
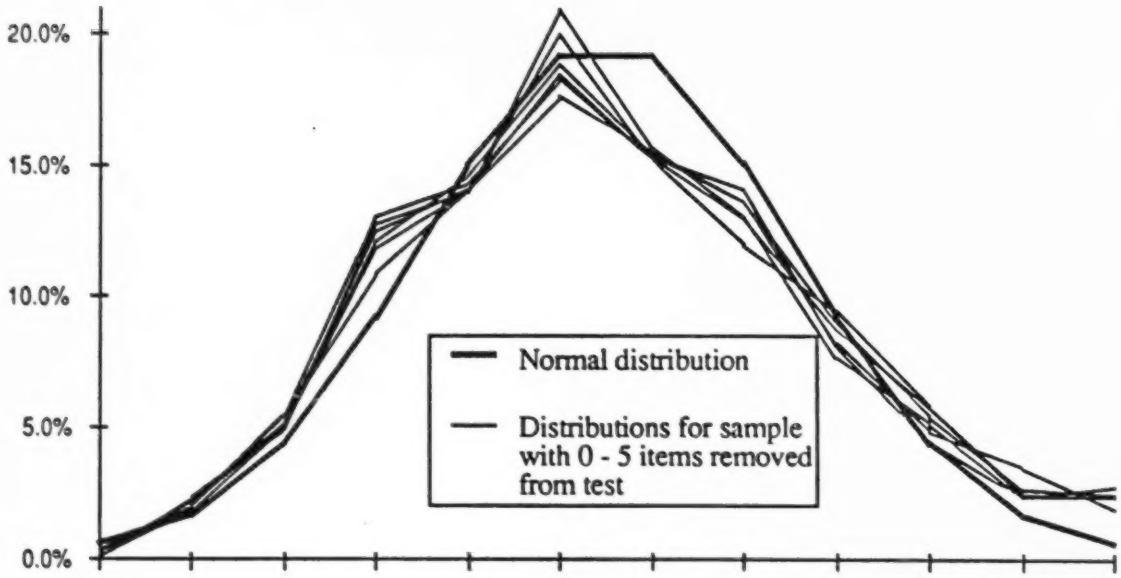


FIGURE 3. DISTRIBUTIONS OF ECI2z FOR SEVENTH-GRADE TESTS

MATH COMPUTATION



READING VOCABULARY

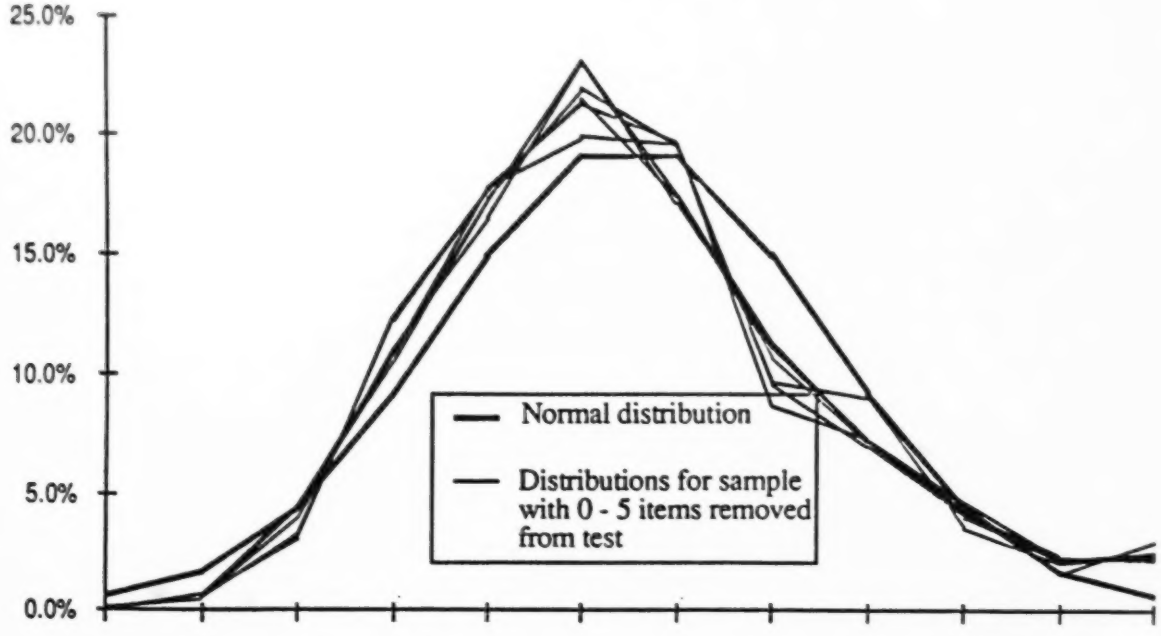
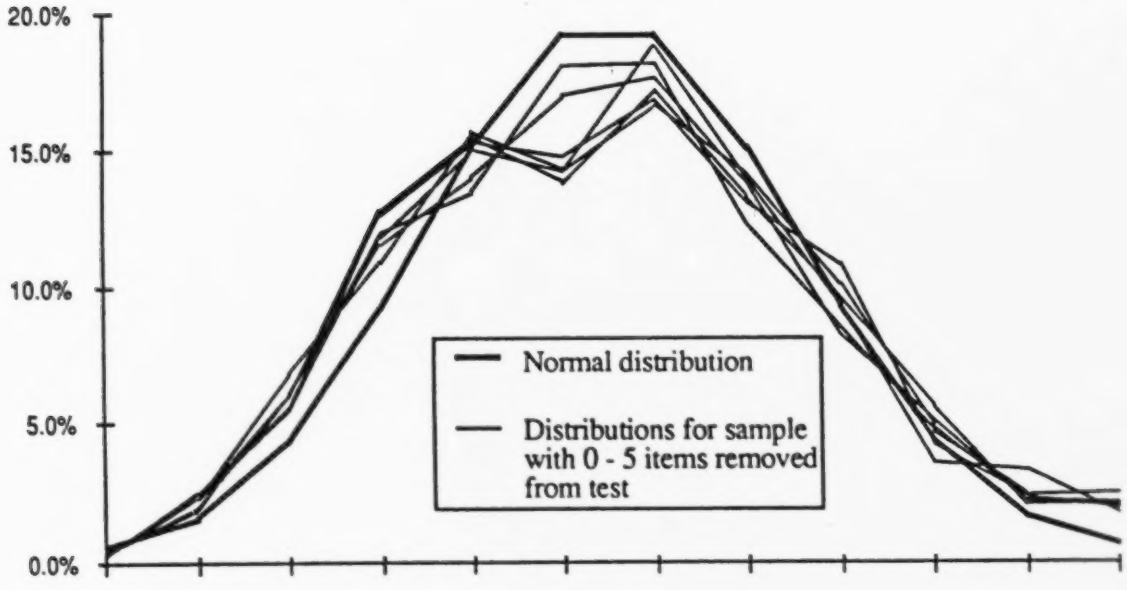


FIGURE 4. DISTRIBUTIONS OF EC14z FOR SEVENTH-GRADE TESTS

MATH COMPUTATION



READING VOCABULARY

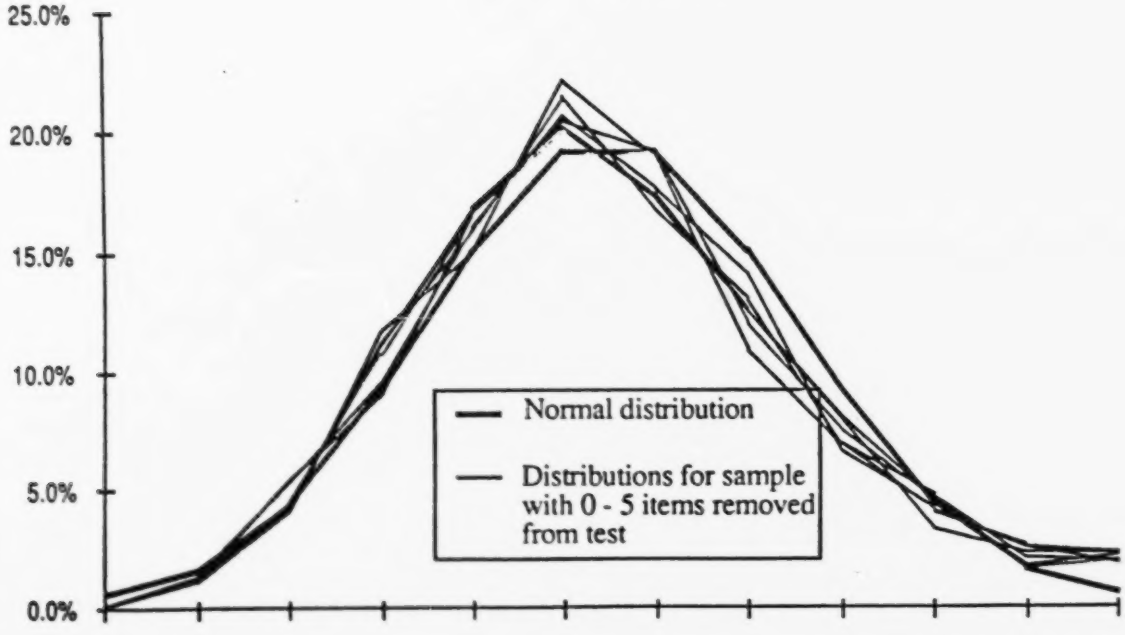


FIGURE 5. SMOOTHED DISTRIBUTIONS OF ECI2z
FOR SEVENTH-GRADE TESTS

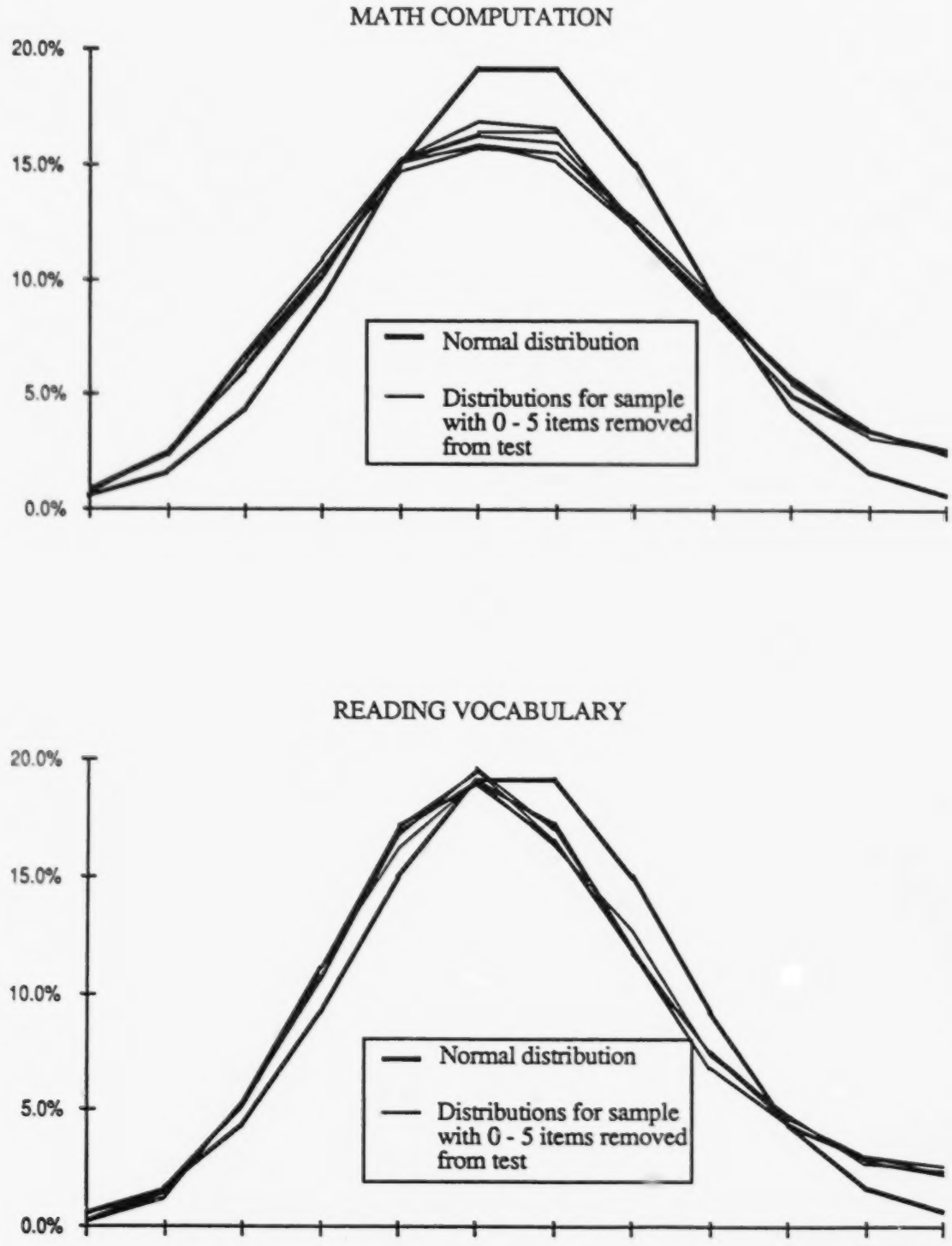
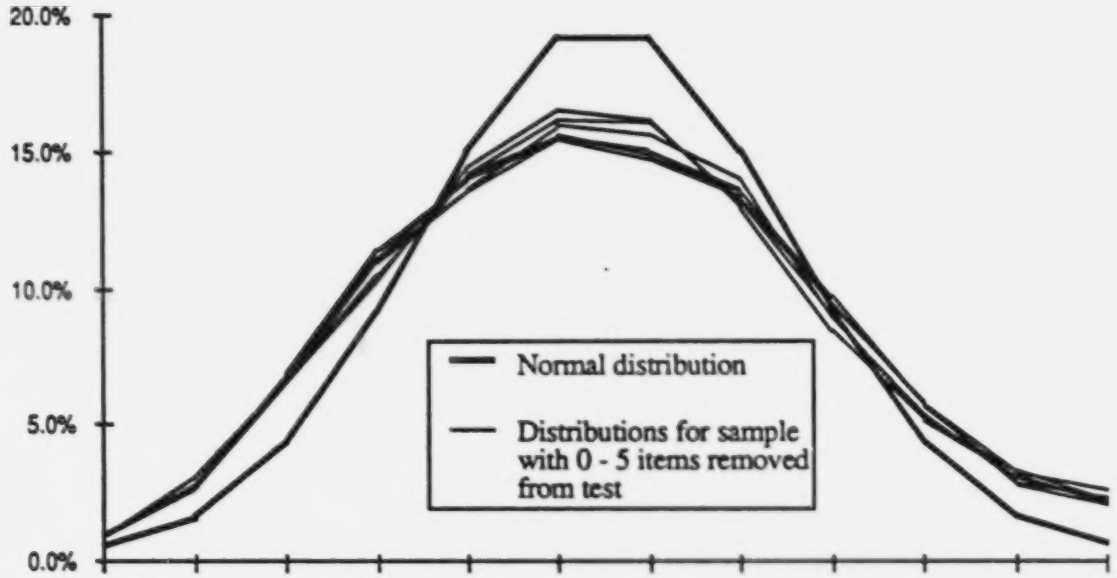
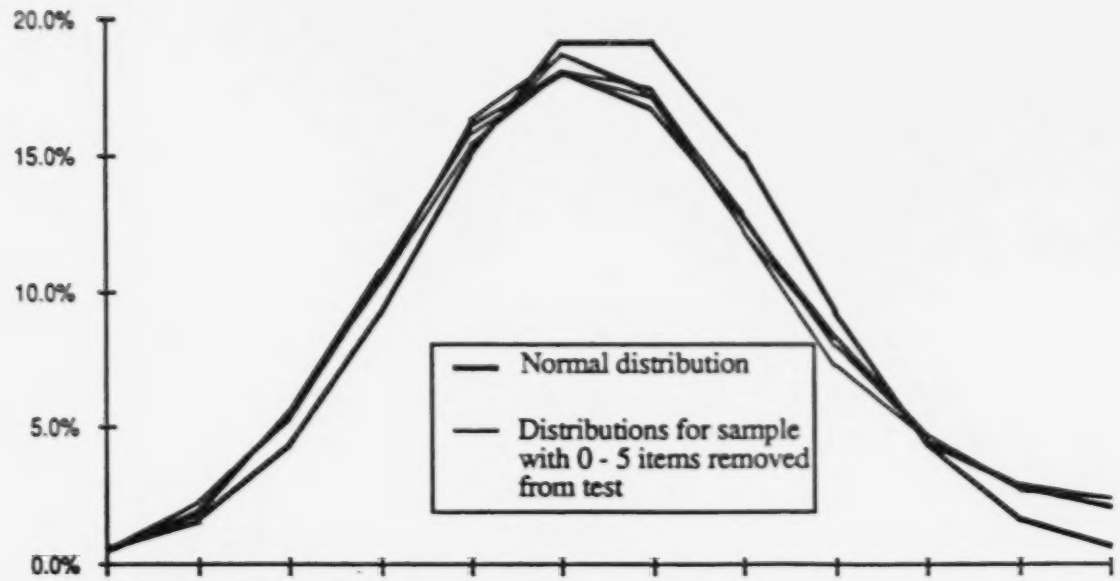


FIGURE 6. SMOOTHED DISTRIBUTIONS OF ECI4z FOR SEVENTH-GRADE TESTS

MATH COMPUTATION



READING VOCABULARY



References

Birenbaum, M. (1985). Comparing the effectiveness of several IRT based appropriateness measures in detecting unusual response patterns. Educational and Psychological Measurement, 45, 523-534.

Divgi, D. R. (1986). Does the Rasch model really work for multiple choice items? Not if you look closely. Journal of Educational Measurement, 23(4), 283-298.

Drasgow, F. (1982). Choice of a test model for appropriateness measurement. Applied Psychological Measurement, 6, 297-308.

Harnisch, D. L. & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. Journal of Educational Measurement, 18(3), 133-146.

Harnisch, D. L. & Tatsuoka, K. K. (1983). A comparison of appropriateness indices based on item response theory. In R. K. Hambleton (Ed.). Applications of item response theory (pp. 104-122). Vancouver: Educational Research Institute of British Columbia.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). Appropriateness measurement. In C. L. Hulin, F. Drasgow, & C. K. Parsons (Eds.) Item Response Theory (pp. 110-151), Homewood, IL.: Dow Jones-Irwin.

Levine, M. V., & Rubin, F. (1979). Measuring the appropriateness of multiple-choice scores. Journal of Educational Statistics, 4, 269-290.

Nelson, R. B., & Chatman, S. P. (1984). In Chatman, S. P. (1985). The relationship between response pattern aberrance and course performance in math placement. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Nelson, R. B., & Chatman, S. P. (1986). The influence of guessing on measures of person-fit when using the Rasch model. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Rudner, L. M. (1983). Individual assessment accuracy. Journal of Educational Measurement, 20, 207-219.

Schmitt, A. P., & Crocker, L. (1984). The relationship between test anxiety and person-fit measures. Paper presented at the Annual Meeting of the American Educational Research Association.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. Psychometrika, 49, 95-110.

Tatsuoka, K. K., & Linn, R. L. (1983). Indices for detecting unusual patterns: Links between two general approaches and potential applications. Applied Psychological Measurement, 7, 81-96.

Tatsuoka, K. K., & Tatsuoka, M. M. (1982). Detection of aberrant response patterns and their effect on dimensionality. Journal of Educational Statistics, 7, 215-231.

Tomsic, M. L. (1986). Stability of extended caution indices for standardized public school testing: A longitudinal study. Unpublished doctoral dissertation, Department of Educational Psychology, University of Oregon.

Tomsic, M. L., & Mittman, A. (1986). Stability of caution indices. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Traub, R. E., & Lam, Y. R. (1985). Latent structure and item sampling models for testing. Annual Review of Psychology, 36, 19-48.

Wright, B. D., & Stone, M. H. (1979). Best test design. Chicago: Mesa.